核酸omicsデータのquality control解析

知久季倫[i]　吉田輝彦[ii]　坂本裕美[iii]

Quality control analyses for nucleotide omics data

Suenori Chiku　Teruhiko Yoshida　Hiromi Sakamoto

国立がん研究センター研究所 遺伝医学研究分野で行っている，1. SNPSチップを用いたgenome wide association sutdyのためのquality control (QC)，2. メチレーションチップのQC，3. 次世代シーケンサーデータのQC解析を報告する．

(**キーワード**): QC, Omics, SNP, Methylation, NGS

1　はじめに

近年，分子生物学の世界では様々な「omics」解析が盛んに行われる様になった．特にゲノムやmRNA発現等の核酸分野では，増幅技術の確立，ヒトゲノムの解読，RNAデータベースの充実等の結果，比較的簡単に網羅的なデータの取得が可能になり，様々な大量データが産み出されている．これらのデータ解析では情報解析技術の適用が必要不可欠であるが，

1. 多層データの解析
2. 次世代シーケンサーの出現

等のニーズからバイオインフォマティクスの重要性が極めて高くなった．本報告では，核酸omicsデータの品質管理における情報学適用の例として，国立がん研究センター研究所遺伝医学研究分野 (遺伝医学) における

1. SNPsチップを用いたgenome wide association sutdy (GWAS)
2. Methylationチップ
3. Next generation sequencing (NGS)
   (a) Whole exome sequencing (WES)
   (b) Whole transcriptome sequencing (WTS)
4. 臨床情報

のquality control (QC) 解析を報告する．

[i]サイエンスソリューション部 バイオエンジニアリングチーム シニアマネージャー 物理学博士
[ii]国立がん研究センター 遺伝医学研究分野 副所長
[iii]国立がん研究センター 遺伝医学研究分野 ユニット長

2　Genome wide association sutdyにおける品質管理解析

2.1　Genome wide association sutdy

ゲノムには個人毎に異る箇所が存在している．ヒトが生まれる際に新たに発生する変異もあるが，殆んどは祖先から受け継いだ変異であるため他人と共有しており，多型と呼ばれる．この多型と表現型 (髪の色等の形質や疾患，薬剤応答性) との関連を調べる方法の一つとして，genome wide association sutdy (GWAS) が近年大きな成果を上げている．多くのGWASは目的の表現型を持つ集団と持たない集団での多型の頻度を比較するcase-control studyであるが，遺伝学における連鎖不平衡を利用したsingle nucleotide polymorphism (SNP) マーカーを使った研究 (図1参照) では100万SNPs単位の比較が行われる．そのため些細なことから大量の偽陽性が発生してしまうが知られており，GWASではQC解析及び集団の構造化解析が極めて重視される．

2.2　GenomeStudioによるSNPsチップのQC

2.2.1　SNP graphによる確認

国立がん研究センター遺伝医学研究分野では，主として(米)Illumina社のOmniシリーズのマイクロアレイでタイピング[1]を行っている．簡単に測定原理を述べると，チップ上に測定したい塩基の隣から60塩基程度のプローブを設置しておき，これに検体のDNAをハ

[1]個人毎のSNPsを測定することをtyping (genotyping) と言う．ヒトは両親から1セットずつの染色体を受け継いでおり，SNPの様に多型がある箇所では多型の組み合わせを測定することになる．遺伝子型 (genptype) とはこの多型の組み合わせであるが，測定データから遺伝子型を特定することをcallと言う.SNPは突然変異率が低いため通常2種類であるが，マイクロサテライトの様な多型では複数種類が存在する．この多型の種類を対立因子 (allele) と呼ぶ．言い換えるとgenotypeとはalleleのペアのことである．

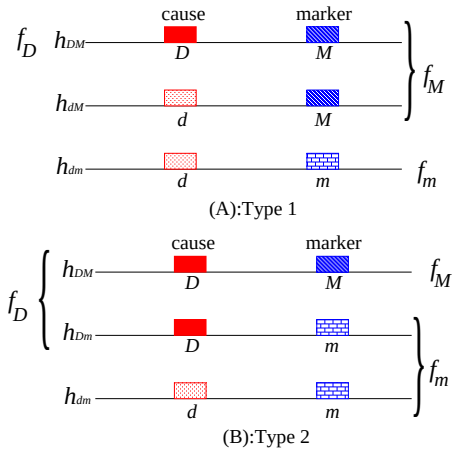

図 1 原因変異 (赤)D と周辺マーカー (青)M との関係の概念図. 3 つ haplotype がある場合を想定 ($D' = 1$). 原因変異と同じ染色体上のマーカーは，世代を重ねることによって染色体の組替えによりやがて独立となるが，近傍では関連が残っていることが期待される (連鎖不平衡). そのため原因変異周辺のマーカーの頻度 f_M も表現型と独立にはならない.

イブリダイゼーションさせ，DNA の伸長反応では二重鎖部分から連続に合成されることを利用して，予め塩基種毎に異る波長の蛍光プローブを付けた塩基で伸長させてこの蛍光を読み取る. 現在の SNPs 用の Omni チップでは，1 枚のチップで 8 ～ 12 検体，1 検体あたり 100 ～ 200 万箇所程度の SNPs を同時測定することが出来る.

Omni チップで測定されたデータは，付属の GenomeStudio と言うソフトウエアで遺伝子型 call が行われる[2]. この時 SNP 毎に SNP graph と呼ばれる図を用いて call 状況を確認することが出来る. 遺伝医学研究分野ではこれを図 2 の様に分類し，GWAS に不適切な SNPs を除いている.

2.2.2 Genome Viewer による確認

検体毎には GenomeStudio の Genome Viewer を利用して品質を確認することができる. Genome Viewer で使用される logR ratio と B allele frequency の定義と表示例を図 3 に示した.SNP graph におけるクラスター中心からの蛍光強度と allele 数の指標を染色体全域について表示させると図 4 に示した様なことを認識することが出来る.

[2]Illumina のチップでは便宜的に A allele と B allele で区別され，AA，AB，BB の何れかで遺伝子型を表す. 尚，AA，BB をホモ接合体 (ホモ)，AB をヘテロ接合体 (ヘテロ) と呼ぶ.

2.3 GWAS のための SNPs チップの dry QC

2.3.1 Dry QC の手順

国立がん研究センター研究所遺伝医学研究分野では，SNPs チップの QC として

1. サンプル毎に call proportion が 98%より低いサンプルの削除
2. ヘテロ割合が高くコンタミネーションと考えられるサンプルの削除
3. 性別が一致しないサンプルの削除
4. 遺伝子型一致率が高いサンプルの削除 (一人のみ残す)
5. 近交係数 F が著しく高いサンプルの削除
6. 構造化解析から日本人 (本土出身者) 以外と考えられるサンプルの削除

を行っている.1. 及び 2. は前述の GenomeStudio からも得られる情報であるが，近年の測定技術の向上により 1. を満たさないサンプルは血液由来では殆んど無くなっている.2. についてはサンプル供給元組織によっては思いの外多いので注意が必要である.

2.3.2 Call proportion と MAF

例としてサンプル毎の call 割合，SNPs 毎の call 割合，minor allele frequency (MAF) のヒストグラムを図 5 に示した. Illumina の Omni チップは主に白人を対象として開発されているため，日本人集団では多型が観察されない SNPs が多数生じてしまう.

2.3.3 性別の確認

X 染色体は男性は 1 本のみであるためヘテロ割合が 0 であることが期待され，女性は Y 染色体上に設計されたプローブは全て欠測値となると期待される (XY 相同領域を除く). 実際に SNPs タイピングの結果から性別や性染色体異常を確認することが出来る. 図 6 に，臨床情報に従って男 (青) 女 (赤) 別に X 染色体のヘテロ割合と Y 染色体の SNP call 割合の散布図を示した. 左上の集団中の赤い三角，右下の集団中の青い丸は臨床情報の間違いである可能性が極めて高い. また右上付近の 2 点はクラインフェルターが疑われ，右下のより右に外れている点は XXX が疑われるサンプルである.

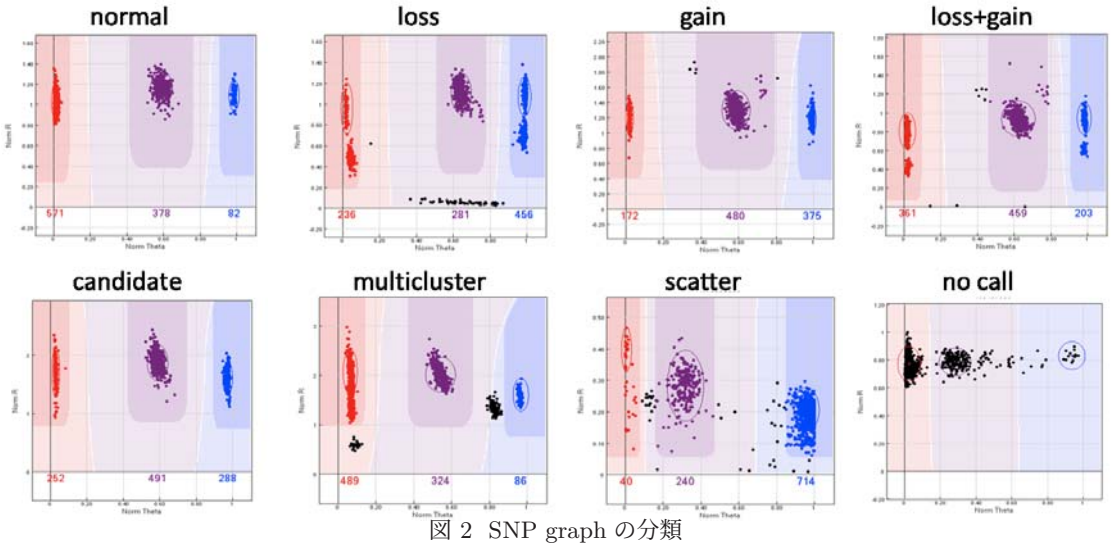

図 2 SNP graph の分類

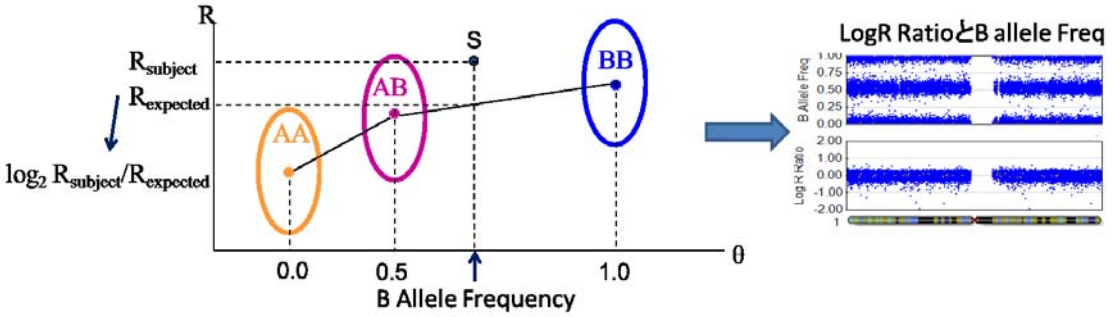

図 3 logR ratio と B allele frequency の定義と Genome Viewer で染色体 1 番を表示した例

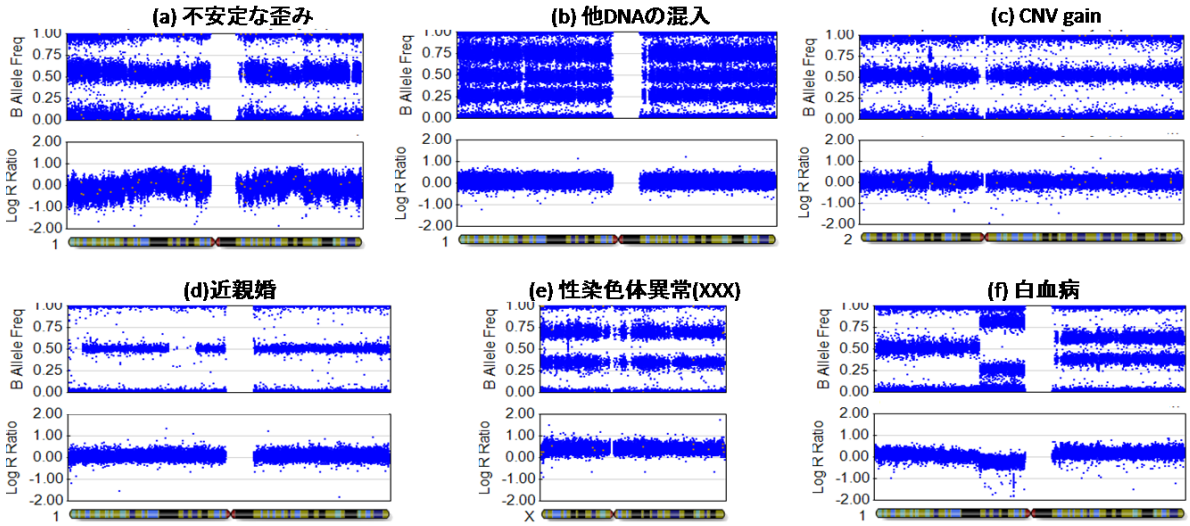

図 4 Genome Viewer により確認できる項目例

2.3.4 遺伝子型一致割合

MAF が 0.5 付近の SNPs を用いると，遺伝子型の一致割合は血縁無し:3/8，親子:1/2，きょうだい:19/32，一卵性双生児若しくは同一人物:1 となることが示される．図 7 の左図に Omni データから計算された例を示す．同じサンプルが異なるサンプルとして提出されることもある．

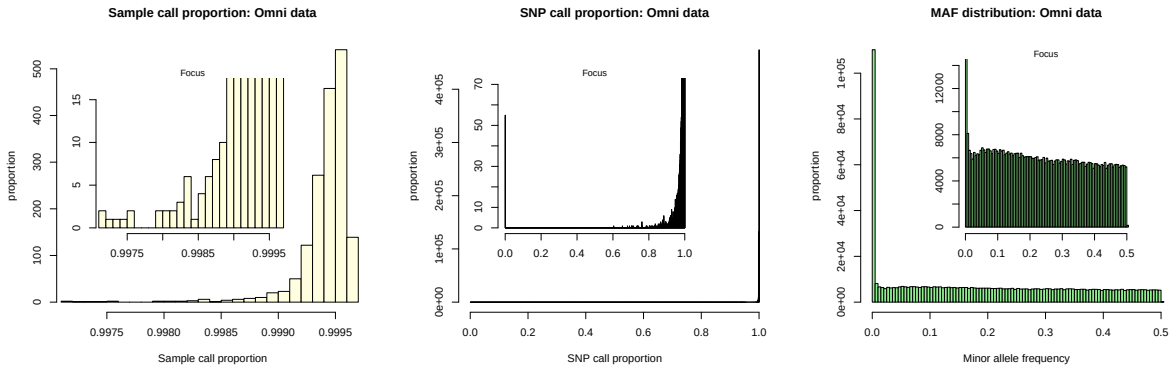

図 5 左図:サンプル毎の call 割合のヒストグラム，中央図:SNP 毎の call 割合のヒストグラム，右図:MAF のヒストグラム

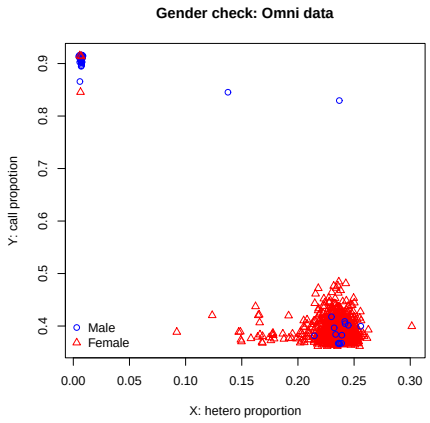

図 6 横軸に X 染色体のヘテロ割合，縦軸に Y 染色体の SNP call 割合とした散布図. 青丸:臨床情報が男性，赤三角:臨床情報が女性.

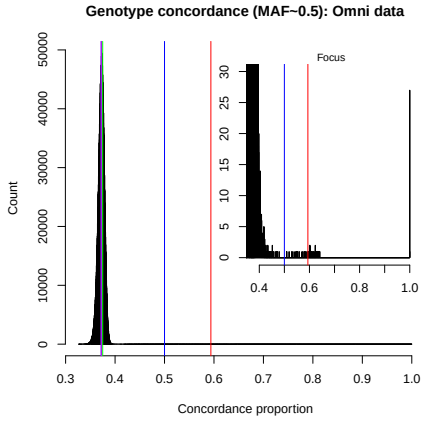

図 7 遺伝子型一致率のヒストグラム.MAF が 0.5 付近の SNPs のみで計算. 赤の縦線はきょうだい，青の縦線は親子，緑の縦線は血縁なしの期待値，紫の縦線は実データの平均一致割合を示す.

2.3.5 近交係数と遺伝子型対数尤度分布

近交係数 F とは育種等の分野で使われている近親婚の指標である．Hardy-Weinberg 平衡 (HWE) を仮定すると allele 頻度が f である SNP ではヘテロと観測される割合は $2f(1-f)$ となるが，近親婚の影響による歪みを $2f(1-f)(1-F)$ と表現することで定義される．よって allele 頻度によって SNP を分類して観測されたヘテロ割合を $2f(1-f)$ で割ることにより $1-F$ を推定することが出来る．F は，いとこ婚，おじめい · おばおい婚，親子 · きょうだい婚でそれぞれ 1/16，1/8，1/4 となる．尚，ここでは集団としての近交係数 F_A を無視している．図 8 にサンプル毎に MAF が 0.01 未満の SNPs を持っている数 (rare allele count) と F の推定値による散布図を示した. 左上の水色の点は親子 · きょうだい婚が疑われるサンプルで，右下の黒点は外国人との混血が疑われるサンプルである.

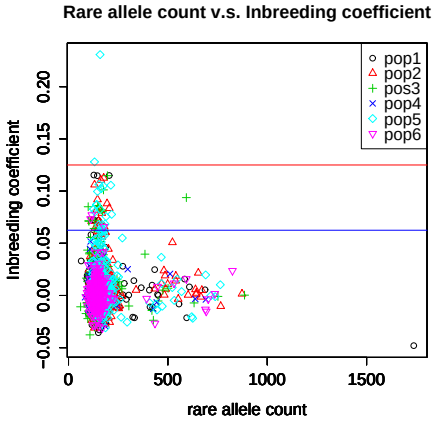

図 8 Rare allele のカウント数と近交係数の散布図. 横の赤線はおじめい · おばおい婚相当，青線はいとこ婚相当を示す.

2.3.6 Hardy-Weinberg 平衡検定

Random mating を仮定すると，allele 頻度が f である SNP の遺伝子型頻度は f^2:$2f(1-f)$:$(1-f)^2$ となることが期待される.HWE 検定ではこの期待値を帰

無仮説として統計的仮説検定を行い p 値を求める．一般に統計的仮説検定は帰無仮説の元で outcome を区間 [0,1] の実数に射影する方法論であるため，GWAS の分野では $-\log_{10} p$ の期待値を横軸に，観測された $-\log_{10} p$ を縦軸にした散布図 (Q-Q plot) を良く作成する．この Q-Q plot により，より興味がある $p < 10^{-7}$ の様な p 値の振る舞いを強調して表示することが出来る (統計的多重比較により Bonferroni 補正を行うと $\alpha = 0.05/10^6 = 5.0 \times 10^{-8}$ となる).Q-Q plot は本来 quantile-quantile plot であるため，p-p plot と呼ばれることもあるが，GWAS の分野では慣例として $(-\log_{10} p)$-$(-\log_{10} p)$ plot を Q-Q plot と呼ぶ．

図 9 に GenomeStudio から出力された全データと SNP graph から call が疑わしい SNPs を除いた場合の HWE 検定 (exact) の Q-Q plot を示した．QC の一つの大きな目標として，HWE 検定の Q-Q plot が y=x の直線に載るように call 不良の SNPs を除くことがある．

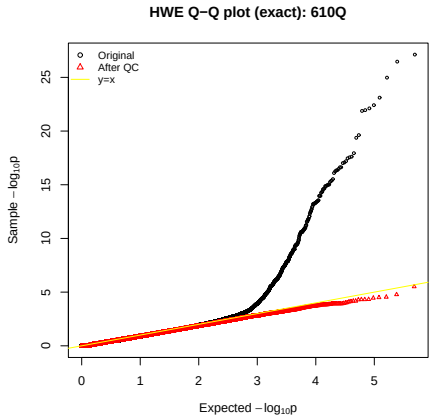

図 9 Hardy-Weinberg 平衡検定 (exact) の Q-Q plot．黒線:全てのデータ，赤線:SNP graph から call が疑わしい SNPs を除いた後，黄色線:y=x.

2.4 集団の構造化解析

2.4.1 集団の構造化解析の目的

GWAS における case-control study では，100 万 SNPs 程度の検定を行うことが多い．そのため人種や地域特異的な特徴を持つ表現型[3]を対象とした場合，人種や地域間での SNPs 頻度の違いを大量に検出してしまう恐れがある．集団の構造化解析では，殆んどの SNPs について帰無仮説が成立していることの確認，若しくは帰無仮説が成立するような control 集団の選択を行う．

2.4.2 Eigenstrat[5] (主成分分析)

Eigenstrat[5] は適当に正規化された遺伝子型データの個人間の揺らぎを主成分分析し，集団の構造化を補正して検定を行う方法論である．実際には主成分分析をした結果を図示することにより，構造化や混血者の検出に用いられることが多い．

Eigenstrat[5] による主成分分析では，国際 HapMap プロジェクト [1] より提供されている白人，黒人，東アジア人等の SNPs データを含めて解析することにより，他人種や混血等を特定することが出来る．図 10 に解析例を示す．図 10 の上段左図からは case 集団中に白人が混ざっていたが分かり (白人集団の側にある赤点)，上段右図からはアジア人においても漢民族と日本人は異るクラスターとして検出可能であることが分かる．ここで上段右図の左側の小集団は沖縄出身者である．下段の図からは北関東集団は沖縄側，関西集団は漢民族側にややずれている様子が見られる．尚，日本人の構造化については文献 [7] に詳しいので参照されたい．

2.4.3 Control 集団の選択

Case-control study における偽陽性の量の指標として，遺伝子型の Cochran-Armitage trend test (CATT) の χ^2 値の中央値を期待値で割った値 $\hat{\lambda}$ (genomic control[2] (GC) と呼ばれる) が慣例的に使われている．遺伝医学研究分野ではこれに加えて，CATT の $-\log_{10} p$ 値を使った Q-Q plot において非有意側 90%の点に対する直線回帰の傾き a，Eigenstrat[5] による p 値の補正の 2 乗平均の平方根 σ も指標に加えている．図 10 の case 集団に対して，control 集団毎の CATT の $-\log_{10} p$ 値の Q-Q plot を図 11 に，$\hat{\lambda}$，a，σ の指標を表 1 に示した．このデータの場合では，検出力も考慮して control 集団として東京+北関東 (T+N) を選択した．

3 Methylation チップにおける品質管理解析

3.1 DNA methylation

ゲノム中には CpG island と呼ばれる CG/GC の頻度が高い領域が存在し，この領域の C がメチル化されると下流の遺伝子発現が抑制されることが知られている．ゲノムのメチル化は epigenetics と呼ばれる研究領

[3] 当初の GWAS では common disease と関連するゲノム領域を特定することを目的とした研究が多かったが，近年は稀な表現型と関連する変異の同定でも成果を上げている．

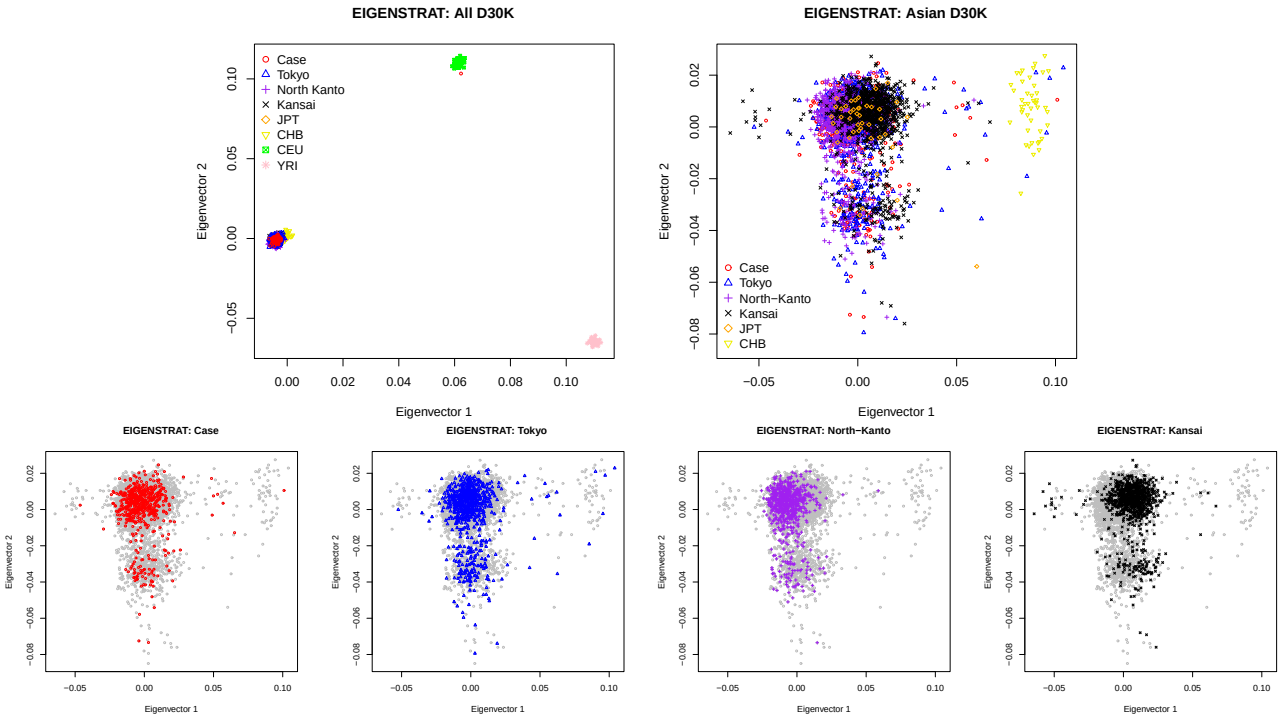

図 10 Eigenstrat[5] による解析結果．上段左図:白人，黒人，東アジア人のデータを含めた解析，上段右図:漢民族と日本人集団．下段は上段右図を集団別に表示．左から case 集団，東京集団，北関東集団，関西集団．

表 1 Case 集団と control 集団の比較における GC $\hat{\lambda}(\chi^2)$，非有意側側 90%データのみにおける $-\log_{10} p(\text{exact})$ の Q-Q plot の傾き a，EIGENSTRAT による trend test (χ^2) の $-\log_{10} p$ 値の補正の 2 乗平均の平方根 σ.T:東京集団，N:北関東集団，K:関西集団，JPT:HapMap の東京集団，CHB:HapMap の漢民族．

statics	東京	北関東	関西	T+N	T+K	N+K	T+N+K	JPT	CHB
$\hat{\lambda}$	1.02	1.13	1.17	1.05	1.11	1.09	1.05	0.952	1.97
a	0.990	1.08	1.13	1.02	1.08	1.05	1.04	0.923	1.66
σ	0.120	0.0927	0.284	0.0480	0.219	0.130	0.130	0.0374	0.906

域の主要なテーマの一つである．Methylation チップでは，DNA を bisulfite 処理して PCR で増幅させると C 塩基が T 塩基に変化するが，メチル化されている C 塩基は変化しないことを利用して，多くの CpG island のメチル化を一度に測定するオミックス技術である．Infinium 社の Human Methylation 450 BeadChip (450K) では 1 枚のチップで 12 検体の 45 万箇所のメチル化を同時に測定することが出来，この 45 万箇所のプローブは CpG island，N_Shelf，N_Shore，S_Shelf，S_Shore，Other (gene body) と分類されている．

3.2 Infinium Human Methylation 450 BeadChip の QC

3.2.1 プローブの QC

Infinium 社のメチル化チップによる測定では，メチル化割合 β は T を検出するプローブと C を検出するプローブの蛍光強度割合 $(0 < \beta \leq 1)$ で定義されている．蛍光強度はバックグランドに対して有意に蛍光が検出されたかどうかの指標である Detection Pval によって足切りすることが出来，遺伝医学研究分野では 0.01 以下のプローブのみを採用している．ここでプローブ毎の call proportion が 0.9 未満となったプローブについては，解析対象から除外している．図 12 に末梢血のメチル化について，常染色体上の各プローブの β 値 call 割合とサンプル毎の call 割合の例を示した．

450K チップの β 値の特徴を示す例として，プローブ

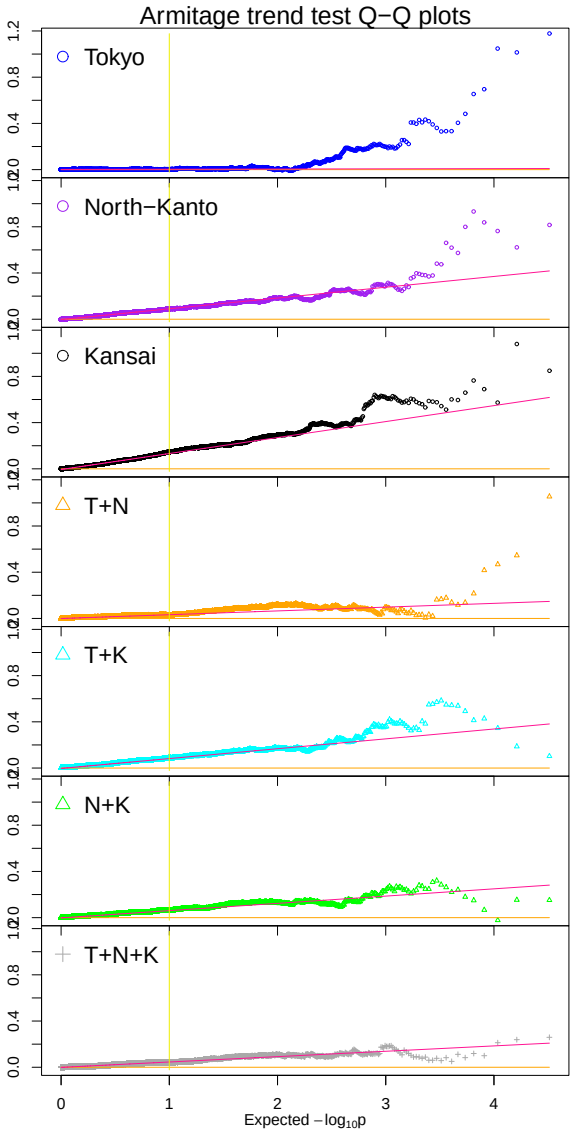

図 11 Case 集団対各 control 集団の組み合わせとの遺伝子型の Cochran-Armitage trend test の $-\log_{10} p$ 値による Q-Q plot. 縦軸は期待値との差. T:東京集団, N:北関東集団, K:関西集団.

カテゴリー毎に常染色体上の平均と分散の散布図及び平均値のヒストグラムを図 13 に示した.Illumina 社から提供されているプローブカテゴリー毎に特徴が異っている.

3.2.2 性別の確認

SNPs チップと同様にメチル化チップからも性別の確認と性染色体異常を検出することが出来る. 女性は X 染色体を 2 本持っているために片方が大域的にメチル化されて不活性化されており, Y 染色体上に設計されたプローブは call されないことが期待される. 図 14 に横軸に X 染色体の全 β 値の平均, 縦軸に Y 染色体

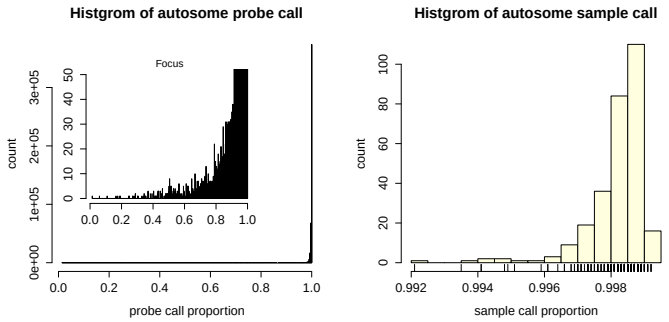

図 12 左図:常染色体上の各プローブの call 割合, 右図:各サンプルの call 割合のヒストグラム

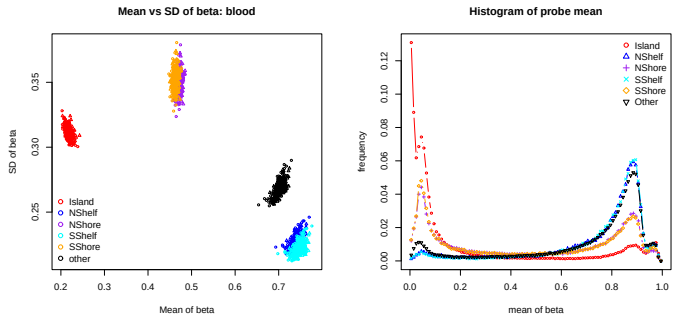

図 13 左図:プローブカテゴリー毎の常染色体上の β 値の平均と分散の散布図, 右図:カテゴリー毎の各プローブの平均 β 値のヒストグラム

の call 割合の散布図を示した. 右下の赤三角の集団中の黒丸は臨床情報と性別が一致しなかったサンプルであり, 右上の黒丸は恐らくクラインフェルターである. 一方で XXX についてはメチル化チップでは検出出来ない様である.

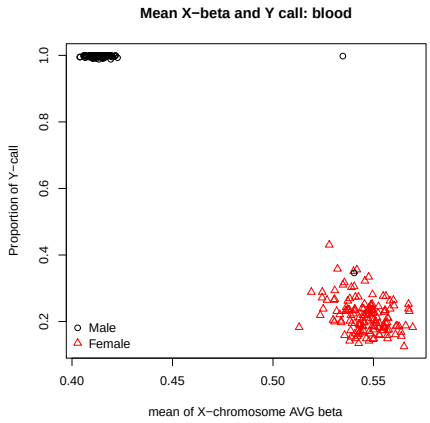

図 14 X 染色体の β 値の平均と Y 染色体の call 割合の散布図.

3.2.3 主成分分析によるバイアス探索

β 値は SNPs と異り区間 $(0, 1]$ の連続値として定義されるため，SNPs チップと比べるとチップのロットや実験日等のバイアスを受けやすい[4]．図 15 に末梢血のメチル化について主成分分析した結果を示した．実験の順序が年齢カテゴリーと交絡している様子が見られる．メチル化は年齢と共に変化することが知られているが，実際に β 値と年齢の重回帰を行うと実験日を調整因子に加えることにより p 値の分布が大きく異ることが図 16 から確認出来る．定量的な omics の場合，実験順序を予めランダム化しておくことが望ましい．

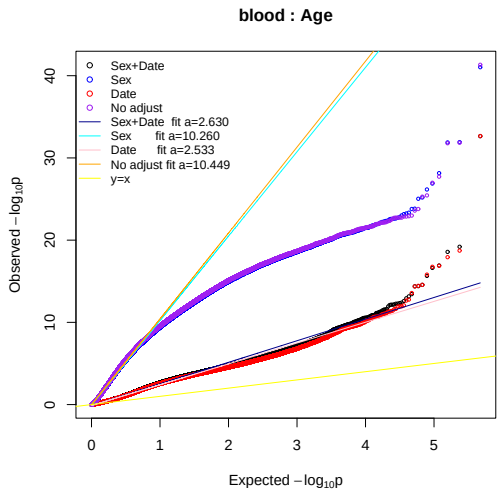

図 16 β 値による年齢に対する重回帰における β 値の $-\log_{10} p$ の Q-Q plot．黒:性別と実験日で調整，青:性別で調整，赤:実験日で調整，紫:調整無し．直線は非有意側側 90%データのみの $-\log_{10} p$ を直線で fitting した結果．

4 Next generation sequencing データの品質管理解析

4.1 Next generation sequencing (NGS)

2010 年頃から 150～500 塩基程度の短い DNA 断片 (insert) の一部の領域 (断片の片側のみや両端 100 塩基程度で read と呼ばれる) を大量 (数十億 read) に解読することが出来るシーケンサーが複数メーカーから販売され，次世代シーケンサーと総称されている．次世代シーケンサーを使った解析では，対象となる核酸の種類によって，

- Whole genome sequencing (WGS)
- Whole exome sequencing (WES)
- Whole transcriptome sequencing (WTS)
- Bisulfite sequencing (BS-seq)
- Chromatin immunoprecipitation sequencing (ChIP-seq)

等がある．

4.2 NGS の QC

4.2.1 QC の項目

遺伝医学研究分野では，主に Illumina HiSeq 2000 を使って WES 及び WTS を行っている.HiSeq ではフローセルと呼ばれるガラス板に 8 本の溝 (lane) が掘られており，更に一つの lane は光学系での蛍光の読み取りにおいて tile と呼ばれる区切り (大まかな物理位置) で分けられている.HiSeq における QC では lane や tile 毎に様々な統計量を計算することで判定を行っている．

NGS のラン毎に作成している要約統計量の例を，表 2-4 に WTS サンプルを GAIIx でシーケンシングした時，表 5-7 に WES サンプルを HiSeq 2000 でシーケンシングした時として示した．これらは表示の問題のため 3 つの表に分かれており，最初の表への記載項目は，

1. サンプル ID
2. lane 番号
3. RIN 値 (RNA の場合)
4. Library の収量 (μg/ml)
5. フローセル上に認識された総クラスター数
6. Illumina のパスフィルターを通過したクラスター数
7. パスフィルターを通過した割合 (%)
8. 完全一致した paired-read 数
9. 完全一致した paired-read 数の割合

である．ここで完全一致した paired-read とは，PCR duplicates と考えている read の一部である (quality が高い場合は殆んど全ての PCR duplicates と考えられる)．これは N (塩基の判定が出来なった場合 N と出力される) や error call も含めての完全一致配列の探索は非常に高速に実施出来るが，N や error call を考慮

[4]SNPs チップも aCGH として CNV 検出等を行う場合には同様である

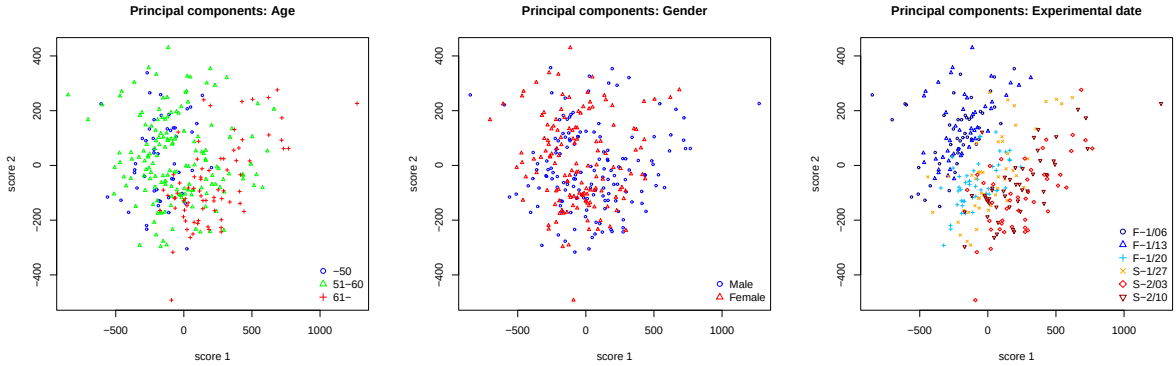

図 15 β 値の相関係数行列に対して主成分分析を行った結果．左から年齢，性，実験日で色分け．

して PCR duplicates を除くにはマッピングを行う必要があり，且つ対象が cDNA の場合にはリファレンス cDNA 配列がまだまだ不十分と考えているために設けている項目である．次に 2 番目の表への記載項目は，

1. サンプル ID
2. 完全一致 paired-read を除いた paired-read 数
3. N，A，T，G，C と call された総塩基数の割合 (%)
4. 全塩基の平均 Q-value (N を含む)
5. A，T，G，C 毎の平均 Q-value

である.Q-value とは読み取った塩基の品質を表す指標 (Phred score) であり，Illumina では区間 $[0, 41]$ の整数で，大きい程エラーが少ないことを表している (ただし N の Q-value は全て 2). 最後に 3 番目の表への記載項目は，

1. サンプル ID
2. BWA[3] で paired-read が uniq (U) 若しくは multi-map (R:repeat) としてアライメント (suffix array 探索でマップ) された割合 (%)
3. BWA[3] で一方の read が Smith-Waterman (M:mate-pair) でアライメントされた割合 (%)
4. Suffix array 探索でマップ (U|R/U|R) された read の edit 数の割合 (%)
5. Suffix array 探索でマップ (U|R/U|R) された read の mismatch 数の割合 (%)
6. Smith-Waterman でマップ (U|R/M) された read の edit 数の割合 (%)
7. Smith-Waterman でマップ (U|R/M) された read の mismatch 数の割合 (%)
8. RNA 及び DNA へのマッピングで SAM 形式のフラグ下 3 つが 011 となり，且つ read 長が insert 長よりも長い paired-read 数 (PPDT). ただし WTS の場合にゲノム上において exon-exon junction に上の read は除いている
9. 上記に対して samtools で potential PCR duplicates (PPD) として除かれた paired-read 数
10. PPD 数 $\times 100$/PPDT 数
11. 上記 PPD を除いた paired-read 数に，proper では無いが RNA 若しくは DNA にマッピングされた paired-read 数を加えた数

である.WTS の場合には UCSC[4] からダウンロードした knownGene，refGene，ensGene を統合した cDNA リファレンス配列にマップし，proper にマップされなかった read を hg19 ゲノムに BWA[3] でマップしている．またこれらの表には示していないが，X 染色体や Y 染色体にマッピングされた read 数の割合や *XIST* 遺伝子 (noncoding RNA) の発現量などで性別の確認を行うことも可能である．

次にサンプル毎に作成している WTS，WES の QC figure の例をそれぞれ図 17，図 18 に示した．これらの図は左上から右向きに順番に，

1. 完全一致 paired-end の数のヒストグラム (log-log plot)
2. BWA[3] による insert length (U|R/U|R と U|R/M 別)
3. Tile 毎の read 数と完全一致 paired-end 削除後の read 数

4. 各 read に含まれる N の数のヒストグラム (read-1:赤線と read-2:青線) と tile 毎の各 read に含まれる N の数のヒストグラム (read-1:ピンク線と read-2:水色線)
5. Cycle 数毎の N の割合 (左縦軸) と各塩基の割合 (右縦軸)
6. 各 tile における cycle 数毎の N の割合
7. 各 tile の N の割合 (左縦軸) と各塩基の割合 (右縦軸)
8. 各 read に含まれる N の数 (左縦軸) と各塩基の数 (右縦軸) のヒストグラム
9. Cycle 数毎の平均 Q-value
10. Tile 毎の各 cycle における平均 Q-value
11. 各 tile の平均 Q-value
12. 各 read 毎の平均 Q-value のヒストグラム
13. BWA[3] でマッピングした結果 (U|R/U|R) の各 cycle におけるミスマッチの割合
14. BWA[3] でマッピングした結果 (U|R/U|R) の tile 毎の各 cycle におけるミスマッチの割合
15. 各 tile のミスマッチの割合 (U|R/U|R)
16. 各 tile の U|R/U|R とマッピングされた割合 (左縦軸) と U|R/M でマッピングされた割合 (右縦軸)
17. BWA[3] でマッピングした結果 (U|R/M) の各 cycle におけるミスマッチの割合
18. BWA[3] でマッピングした結果 (U|R/M) の tile 毎の各 cycle におけるミスマッチの割合
19. 各 tile のミスマッチの割合 (U|R/M)
20. 各 tile の edit distance と mismatch の割合. 左縦軸:U|R/U|R，右縦軸:U|R/M
21. WTS の場合，MLL2，DYNC1H1，SORL1，KIF1C，PRPF8 遺伝子にマップされた read のカウント数 (depth:黒線). 縦の黄色線は exon の境界を示し，赤線は黒線をスムージングした結果 (青の横線は RNA の 5' 末端付近，中間点付近，3' 末端付近の 400bp の平均 depth を表しており，これらの比を図中に示している)
22. WGS 及び WES の場合，bait 領域及び target 領域における depth の分布とその拡大

である. ここで bait 領域とは (米)Agilent Technologies 社の sure select 製品 (WES で用いる exon 領域の DNA 濃度を上げるキット) で設計されている短い DNA 断片長の集合であり，target 領域とは bait 領域に ±100bp を加えた領域のことである.

4.2.2 QC の case study と post QC

NGS データの実際の QC の適用例として，特定のタイルで測定精度が下がっていたためこれを削ったケースを図 19 に，read-2 の最後の 5 塩基を削った場合を図 20 に示した.

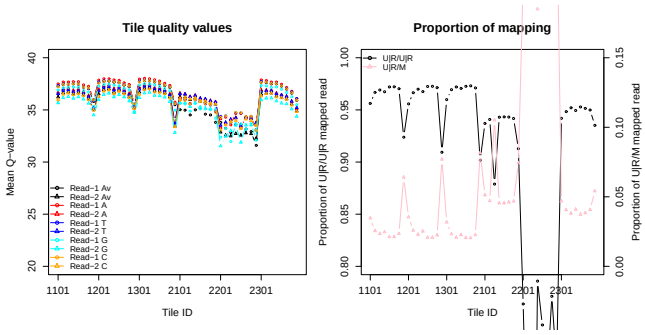

図 19 左図:タイル毎の平均 Q-value，右図:タイル毎のマッピング割合. 2201∼2208 タイルのデータを削った.

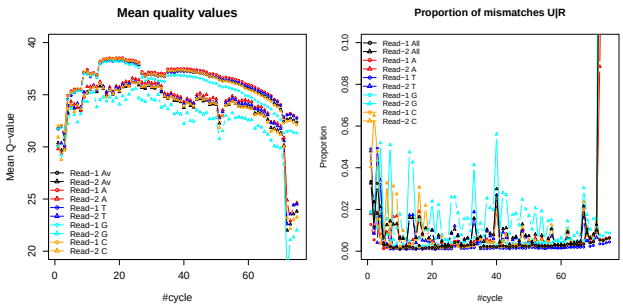

図 20 左図:Cycle 毎の平均 Q-value，右図:cycle 毎のミスマッチ割合 (U|R). Read-2 の最後の 5 塩基を削った.

解析対象データが定量的である場合，様々な実験バイアスが生じることはチップを使ったオミックスデータと同様である. 図 21 に WTS データにおける転写物の発現量推定値である FPKM[6] 値が，主成分分析によって実験施設 (ライブラリー作成)，試薬のバージョン，read 長でクラスターを形成している例を示した.

表 2 120823RNASeq: lane information and number of exactly identical reads

ID	lane#	RIN	Lib(μg/ml)	#cluster	#PF	PF%	#identical	%ident
BR0161DTTS01	1	7.2	23.44	57,563,925	48,034,407	83.4453	4,298,227	8.9482
BR0125DTTS01	2	7.4	20.42	56,655,608	47,662,967	84.1275	4,703,652	9.8686
BR0127DTTS01	3	6.6	18.59	50,262,751	43,736,126	87.0150	5,647,201	12.9120
BR0128DTTS01	4	6.3	22.47	54,097,853	46,351,955	85.6817	4,919,638	10.6137
BR0162DTTS01	5	8.3	23.39	54,573,550	46,781,302	85.7216	5,406,979	11.5580
BR0163DTTS01	6	6.1	26.76	47,721,523	42,086,493	88.1918	4,667,189	11.0895
BR0137DTTS01	7	7.3	25.57	56,094,153	47,313,336	84.3463	5,288,850	11.1783
BR0138DTTS01	8	8.1	22.65	54,564,377	46,001,083	84.3061	4,693,220	10.2024

表 3 120823RNASeq: Base call informations

ID	#paired-read	read	N%	A%	T%	G%	C%	$\bar{Q}$	Q_A	Q_T	Q_G	Q_C
BR0161DTTS01	43,736,180	1	0.0721	28.56	28.50	21.43	21.43	36.78	36.83	36.91	36.49	36.93
		2	0.1389	28.35	28.23	21.28	22.01	35.84	35.99	36.14	35.46	35.86
BR0125DTTS01	42,959,315	1	0.0536	28.10	28.05	21.87	21.93	36.55	36.62	36.71	36.24	36.65
		2	0.1197	27.90	27.88	21.66	22.44	35.28	35.51	35.68	34.80	35.16
BR0127DTTS01	38,088,925	1	0.0475	26.87	26.89	22.96	23.24	36.53	36.50	36.71	36.33	36.61
		2	0.1335	26.82	26.53	23.03	23.49	34.43	34.66	35.07	33.76	34.29
BR0128DTTS01	41,432,317	1	0.0484	27.64	27.63	22.34	22.35	36.41	36.41	36.59	36.13	36.53
		2	0.1321	27.52	27.44	22.10	22.81	35.87	36.04	36.25	35.44	35.81
BR0162DTTS01	41,374,323	1	0.0510	27.79	27.79	22.11	22.26	36.73	36.78	36.90	36.41	36.84
		2	0.1328	27.58	27.62	21.91	22.75	35.65	35.87	36.05	35.19	35.55
BR0163DTTS01	37,419,304	1	0.0509	27.39	27.41	22.47	22.68	37.09	37.08	37.28	36.84	37.21
		2	0.1338	27.28	27.09	22.54	22.96	35.41	35.63	36.00	34.70	35.34
BR0137DTTS01	42,024,486	1	0.0559	27.82	27.80	22.18	22.14	36.57	36.66	36.77	36.20	36.66
		2	0.1496	27.65	27.55	21.89	22.76	35.48	35.76	35.95	34.96	35.30
BR0138DTTS01	41,307,863	1	0.0598	27.03	27.01	22.87	23.03	36.50	36.59	36.74	36.12	36.56
		2	0.1585	26.73	26.86	22.67	23.58	35.27	35.61	35.84	34.71	35.01

表 4 120823RNASeq: BWA mapping informations (hg19 RNA)

	mapping%		U|R/U|R		U|R/M		Effective reads in SAM			
ID	U|R/U|R	U|R/M	edit%	mis%	edit%	mis%	#PPDT	#PPD	%PPD	residue
BR0161DTTS01	71.9792	0.9064	0.3707	0.3441	3.4619	2.9690	41,994,922	2,338,298	5.5680	39,902,230
BR0125DTTS01	74.4049	1.0288	0.3486	0.3219	3.5483	3.0774	40,871,826	1,970,892	4.8221	39,132,557
BR0127DTTS01	70.7423	1.2335	0.3743	0.3492	3.5187	3.1706	34,776,378	2,406,856	6.9210	32,596,082
BR0128DTTS01	69.5765	1.0522	0.3566	0.3279	3.4879	3.0049	39,572,411	1,921,012	4.8544	37,884,272
BR0162DTTS01	75.3799	0.9255	0.3304	0.3048	3.4721	2.9982	39,443,267	2,017,429	5.1148	37,665,883
BR0163DTTS01	65.8511	1.2894	0.3513	0.3241	3.7643	3.3538	34,825,265	1,821,221	5.2296	33,262,969
BR0137DTTS01	68.9449	1.1227	0.3631	0.3346	3.6380	3.1714	39,980,316	2,146,576	5.3691	38,100,815
BR0138DTTS01	69.6242	1.2193	0.3549	0.3310	3.6920	3.2646	38,952,727	1,973,012	5.0651	37,231,756

表 5 120926HiSeqA: lane information and number of exactly identical reads

ID	lane#	RIN	Lib(μg/ml)	#cluster	#PF	PF%	#identical	%ident
BR0127DTGX01	5	NA	??	105,555,644	94,009,411	89.0615	17,243,316	18.3421
BR0128DTGX01	5	NA	??	99,942,427	89,589,043	89.6407	21,549,437	24.0537
BR0101NTGX01	6	NA	??	97,638,282	86,892,694	88.9945	17,878,221	20.5751
BR0107NTGX01	6	NA	??	106,094,321	95,390,051	89.9106	19,348,688	20.2838
BR0108NTGX01	7	NA	??	106,772,902	94,165,512	88.1923	18,773,398	19.9366
BR0112NTGX01	7	NA	??	106,288,196	94,041,750	88.4781	18,038,569	19.1814
BR0119NTGX01	8	NA	??	103,212,819	91,569,605	88.7192	16,059,042	17.5375
BR0125NTGX01	8	NA	??	103,642,452	91,737,394	88.5133	17,295,139	18.8529

表 6 120926HiSeqA: Base call informations

ID	#paired-read	read	N%	A%	T%	G%	C%	$\bar{Q}$	Q_A	Q_T	Q_G	Q_C
BR0127DTGX01	76,766,095	1	0.0020	25.80	25.88	23.52	24.79	35.67	35.85	35.82	35.26	35.70
		2	0.0451	25.92	25.99	24.49	23.56	34.93	35.27	35.17	34.43	34.88
BR0128DTGX01	68,039,606	1	0.0021	25.84	25.89	23.55	24.71	35.65	35.84	35.82	35.23	35.67
		2	0.0498	25.92	26.09	24.39	23.56	34.89	35.25	35.15	34.38	34.82
BR0101NTGX01	69,014,473	1	0.0022	25.84	25.93	23.52	24.71	35.94	36.14	36.08	35.55	35.97
		2	0.0288	25.96	26.04	24.41	23.56	35.26	35.58	35.47	34.80	35.21
BR0107NTGX01	76,041,363	1	0.0022	25.81	25.94	23.47	24.78	36.02	36.22	36.16	35.63	36.04
		2	0.0329	26.04	25.94	24.53	23.45	35.24	35.55	35.44	34.78	35.21
BR0108NTGX01	75,392,114	1	0.0015	25.98	25.87	23.62	24.52	35.68	35.88	35.83	35.27	35.71
		2	0.0406	26.02	26.21	24.21	23.52	35.04	35.35	35.27	34.54	35.00
BR0112NTGX01	76,003,181	1	0.0015	25.61	25.74	23.60	25.06	35.73	35.93	35.88	35.34	35.76
		2	0.0381	25.81	25.73	24.79	23.63	35.05	35.35	35.28	34.57	35.02
BR0119NTGX01	75,510,563	1	0.0019	25.84	25.94	23.47	24.75	35.97	36.17	36.09	35.60	36.00
		2	0.0337	26.01	26.04	24.44	23.47	35.26	35.55	35.46	34.79	35.24
BR0125NTGX01	74,442,255	1	0.0019	26.13	26.27	23.15	24.45	35.98	36.18	36.11	35.59	36.00
		2	0.0345	26.30	26.36	24.10	23.21	35.19	35.48	35.41	34.70	35.17

表 7 120926HiSeqA: BWA mapping informations (hg19)

	mapping%		U|R/U|R		U|R/M		Effective reads in SAM			
ID	U|R/U|R	U|R/M	edit%	mis%	edit%	mis%	#PPDT	#PPD	%PPD	residue
BR0127DTGX01	92.1232	5.5842	0.3446	0.3355	4.5938	4.1303	75,006,082	8,042,436	10.7224	67,764,273
BR0128DTGX01	92.1451	5.7902	0.3576	0.3484	4.6128	4.1463	66,634,743	9,395,923	14.1006	57,784,822
BR0101NTGX01	93.6328	4.8329	0.3259	0.3174	4.5879	4.1083	67,955,578	7,672,417	11.2903	60,673,355
BR0107NTGX01	92.7080	4.6895	0.3140	0.3056	4.5923	4.1144	74,062,323	7,784,071	10.5102	67,439,054
BR0108NTGX01	92.4923	5.0848	0.3449	0.3363	4.6697	4.2226	73,565,491	8,804,131	11.9677	65,603,872
BR0112NTGX01	92.0841	5.1573	0.3367	0.3282	4.6180	4.1706	73,906,502	8,095,807	10.9541	66,883,315
BR0119NTGX01	93.3847	4.6278	0.3125	0.3040	4.6351	4.1465	74,009,810	6,950,403	9.3912	67,749,398
BR0125NTGX01	93.5694	4.7439	0.3230	0.3143	4.6722	4.1856	73,186,563	7,776,584	10.6257	65,861,173

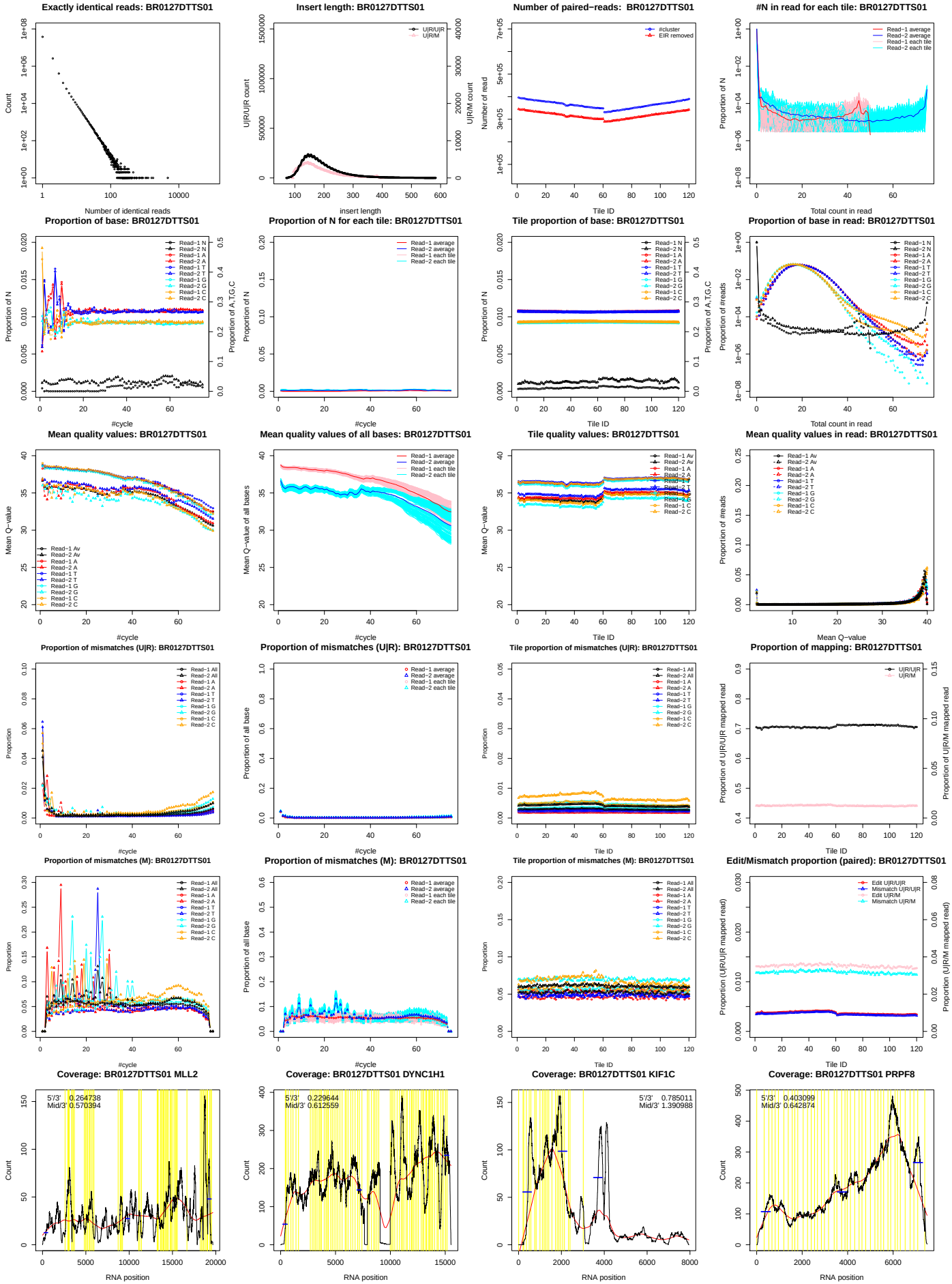

図 17 BR0127DTTS01 (120823RNASeq:lane3) RIN=6.6, Lib=18.59 μg/ml QC figures

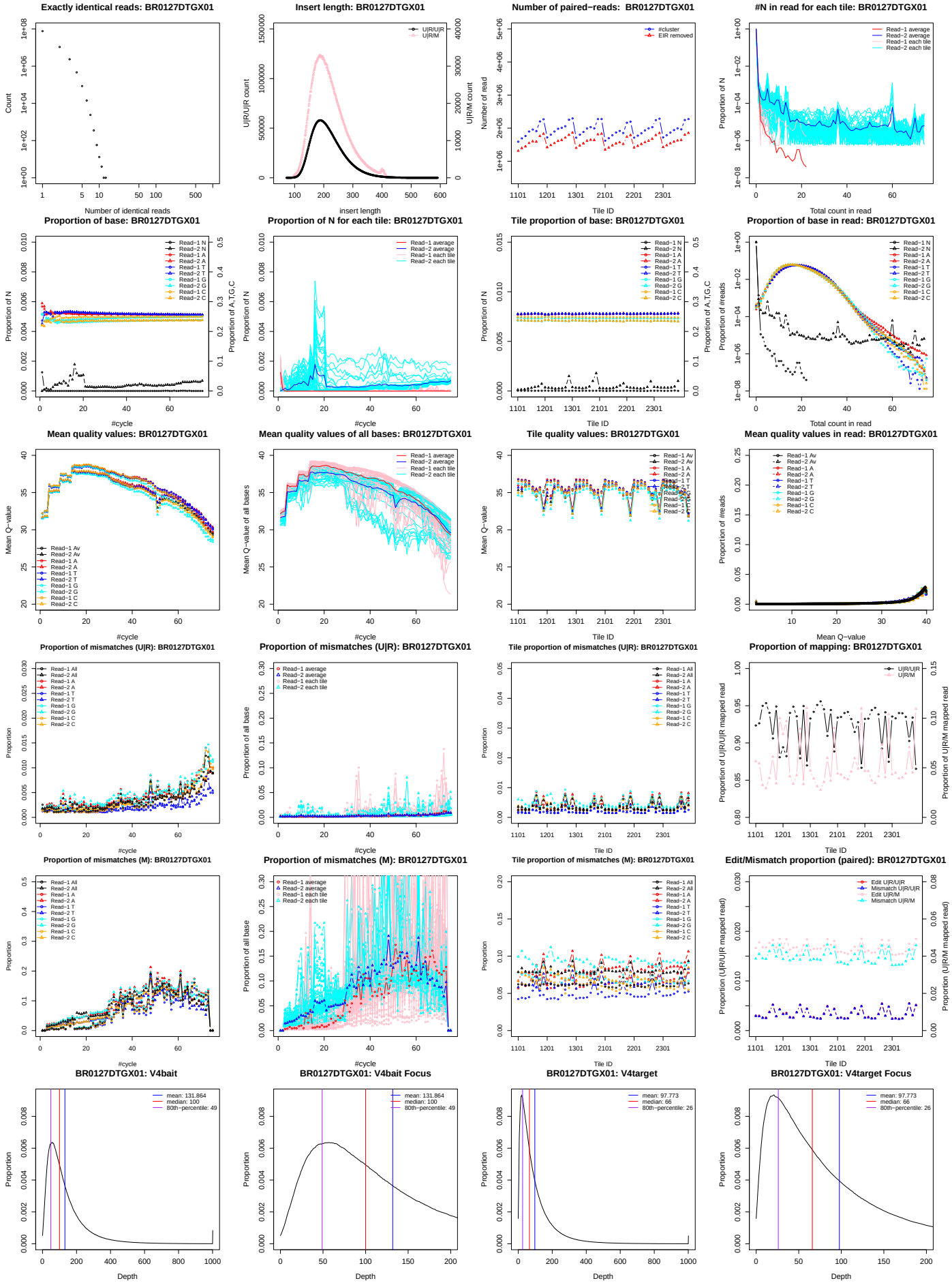

図 18 BR0127DTGX01 (120926HiSeqA:lane5) QC figures

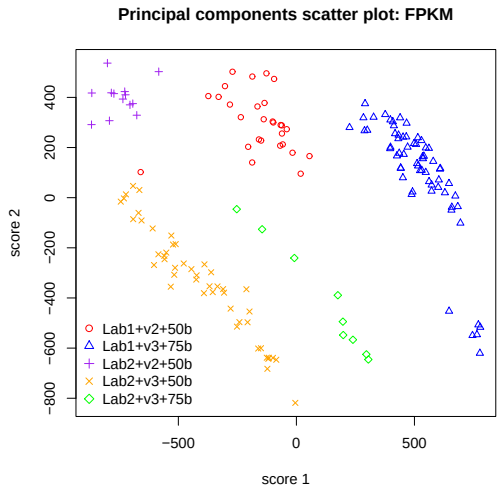

図 21　FPKM[6] 値の分散共分散行列を主成分分析した時のスコア 1 と 2 の散布図.Lab1 と Lab2 は NGS のランを行った施設，v2 及び v3 は試薬のバージョン，50b 及び 75b は read 長. オレンジ色のクラスターにある赤丸は，ライブラリー作成を Lab2 が行っている.

5　臨床情報の品質管理解析

臨床情報は人が直接入力する情報であるため，どんなに注意深く作成しても全くミスが入らないことは極めて困難である. 臨床情報は Microsoft excel 形式で届くことが多く，これまでの経験から情報解析者として

1. 全角英数字，全角スペース
2. 数字やフラグ箇所への自然言語での書き込み
3. 色で区別
4. 表記の揺れ
5. 臨床情報は時系列データであり，論文化までに最低 3 回は全計算をやり直す

らを意識して注意している. また臨床情報と臨床検体の突合わせも重要であり，

1. 経験的に 200 人に 1 人くらいは性別が一致しない
2. 同じサンプルが異る ID で複数回提供される
3. 地域的に異る施設間でも血縁者が含まれる
4. 健常人として提供された血液サンプルが白血病だった
5. 日本人ではなかった
6. がん部，非がん部のペアが別人
7. がん部と非がん部の標識ラベルの付け間違え
8. 他者の DNA の混入

らを QC として確認している. 上記の精度は検体提供施設によって大きな差がある.

謝辞：本稿は国立がん研究センター遺伝医学研究分野において 10 年以上に渡り培ってきた quality control 解析を纏めたものである[5]. 本稿で例として示した図表は，国立がん研究センターの河野隆志分野長，岩崎基部長，及び国立国際医療研究センターの安田和基部長らのデータを使用させて頂いた. ここに慎んで感謝の意を表したい.

引 用 文 献

1) International HapMap Consortium. Integrating ethics and science in the international hapmap project. *Nature Revivew Genetics*, Vol. 5, No. (6), pp. 467–475, Jun 2004.

2) B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, Vol. 55, pp. 997–1004, 1999.

3) Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, Vol. 25, No. (14), pp. 1754–60, Jul 2009.

4) Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, , and Kent WJ. The ucsc genome browser database: extensions and updates 2013. *Nucleic Acids Res*, Vol. 41, No. D1, pp. D64–D69, Nov 2012.

5) A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, Vol. 38, No. 8, pp. 904–909, Aug 2006.

[5] 2012 年に日本臨床試験研究会での発表を契機にまとめた.

6) Cole Trapnell and *et al.* Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*, Vol. 28, No. (5), pp. 511–515, May 2010.

7) Y. Yamaguchi-Kabata, K. Nakazono, A. Takahashi, S. Saito, N. Hosono, M. Kubo, Y. Nakamura, and N. Kamatani. Japanese population structure, based on snp genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am. J. Hum. Genet.*, Vol. 83, No. 4, pp. 445–456, Oct. 2008.